Intelligent Archives
A Conceptual Architecture Study

H. K. Ramapriyan, Steve Kempler, Chris Lynnes, Gail McConaughy, Ken McDonald, Richard Kiang
NASA Goddard Space Flight Center

Sherri Calvo, Robert Harberts, Larry Roelofs,
Global Science and Technology, Inc.

Donglian Sun
George Mason University

Goal

To create a next generation conceptual archive architecture supported by advanced technology that is able to:

- Increase data utilization by hosting and applying IDU technologies such as:
  - Information and knowledge extraction
  - Automated data object identification and classification
  - Intelligent user interfacing, and system management
  - Distributed computing and data storage
- Automate the transformation of data to information and knowledge allowing the user to focus on research/applications rather than data and data system manipulation
- Exploit new and emerging technologies as they become available
- Incorporate lessons learned from existing archives
- Accommodate new data intensive missions without redesign or restructuring

Technical Objectives

- Formulate concepts and architectures that support data archiving for NASA science research and applications in the 10 to 20 year time frame
- Focus on architectural strategies that will support intelligent processes and functions
- Identify and characterize science and applications scenarios that drive intelligent archive requirements
- Assess technologies and research that will be needed for the development of an intelligent archive
- Identify and characterize potential IDU research projects that will be needed to develop and create an intelligent archive

The Problem

- Most of NASA’s archived data is spatial (images) and temporal in nature with minimal information about data content
- NASA’s scientific data holding are becoming voluminous
  - Increasing numbers and kinds of data sources (sensors, users, new missions, etc.) are generating large quantities of data and information
  - Model data volumes are expected to rival remotely sensed data
- Presently image analysis and feature identification can only be successfully performed by human experts
  - Human-based strategies for managing, searching, identifying, and creating required data and information for research purposes are time-consuming and cost-prohibitive for large archives
  - Acquisition and accumulation rates continue to outpace the ability to manage, discover, and exploit scientifically meaningful data, information and knowledge
- Extremely difficult to automate the data, information, & knowledge extraction processes

The Problem (Continued)

- Existing archives neither have the architecture nor technologies to support automated intelligent data understanding
- Archives and service providers are distributed and belong to diverse institutions with their own data organization and access mechanisms
- Contributes to heterogeneous data, information and knowledge
  - Interoperability is a significant driver
- Tools to support automated identification, and classification of objects and events are being developed but must be matched with complementary archive architectures to be successful
- Existing archives suffer from the fact that
  - Every generation tends to use different technologies and architectures that are driven by schedule and cost
  - Software is hardware and application specific

What Is An Intelligent Archive (IA)?

- An IA includes all items stored to support “end-to-end” research and applications scenarios
- Stored items include:
  - Data, information and knowledge
  - Software and processing needed to manage holdings and improve self-knowledge (e.g., data-mining to create robust content-based metadata)
  - Interfaces to algorithms and physical resources to support acquisition of data and their transformation into information and knowledge (could be invoked in push or pull mode)
  - Architecture expected to be highly distributed so that it can easily adapt to include new elements as data and service providers
- Will have evolved functions beyond that of a traditional archive
  - The “borders” of an intelligent archive are intrinsically fuzzy, but may be determined in practice by institutional structure and expectations
  - Will be based on and exploit technologies in the 10 to 20 year time range
- Will be highly adaptable so as to meet the evolving needs of science research and applications in terms of data, information and knowledge

Data, Information and Knowledge

Data: output from a sensor, with little or no interpretation applied

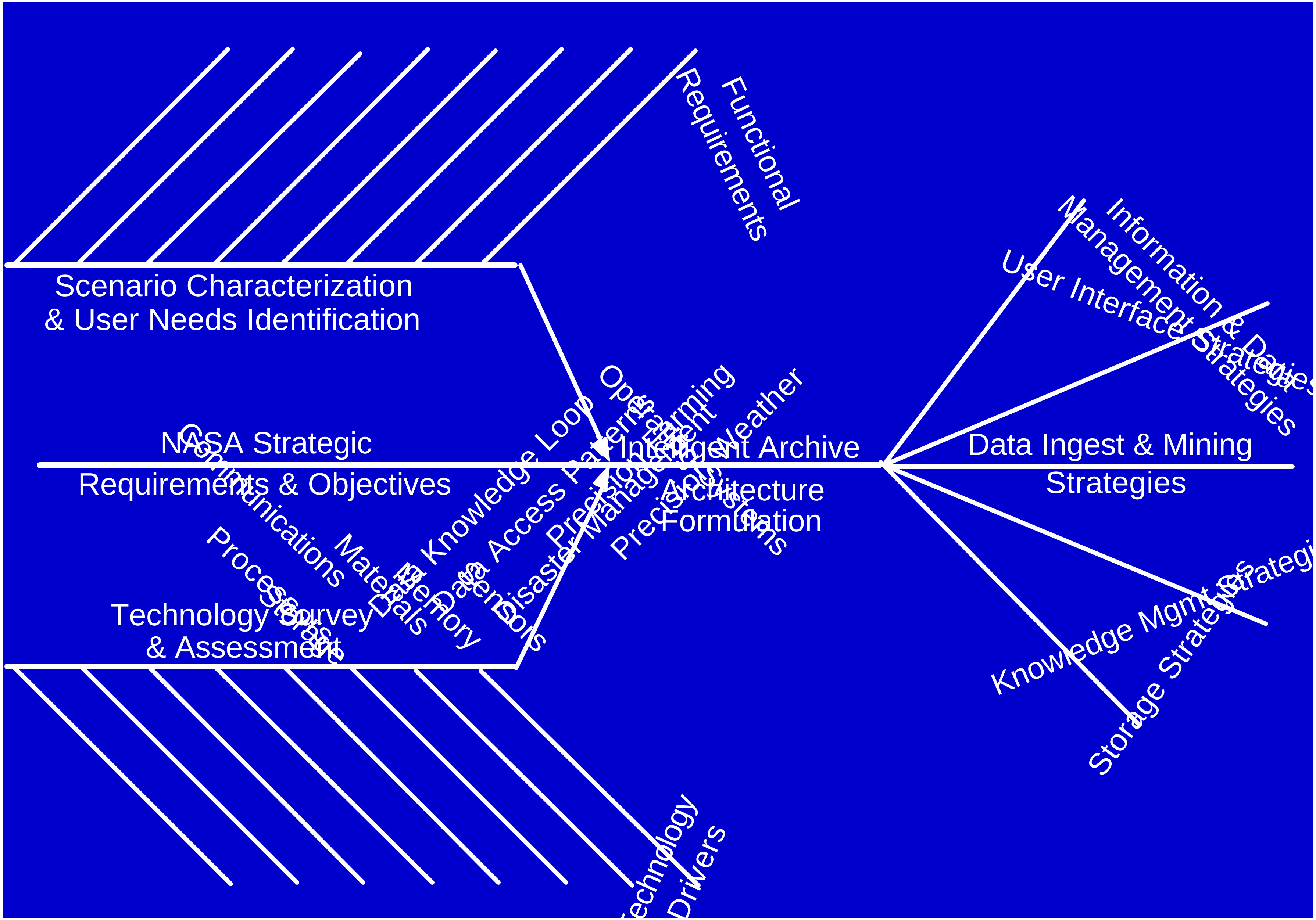
- Examples: *Scientific instrument measurements, market past performance*

Information: a summarization, abstraction or transformation of data that increases our understanding of the physical world

- Examples: *Results after performing transformations by data mining, segmentation, classification, etc., such as a Landsat scene spatially indexed based on content , assigned a "class" value and subset for an application, or National Weather Service storm monitoring fused with a GIS of the spatial location of the Washington D.C. Beltway.*

- Knowledge: a summarization, abstraction or transformation of information that allows our understanding of the physical world
- Examples: *Predictions from model forward runs, published papers, output of heuristics or other techniques applied to information to answer a "what if" question such as "What will the accident rate be if an ice storm hits the Washington D.C. Beltway between Chevy Chase and the Potomac crossing at* 7 a.m.?"

Approach


Functional Requirements
Scenario Characterization & User Needs Identification
NASA Strategic Requirements & Objectives
Technology Survey & Assessment
Intelligent Archive Architecture Formulation
Information & Data Management Strategies
User Interface Strategies
Data Ingest & Mining Strategies
Knowledge Mgmt Strategies
Storage Strategies
Technology Drivers
Communications
Materials
Data Knowledge Loop
Data Access Patterns
Disaster Management

NASA Relevance

- Earth Science has large archive holdings that are growing at an ever increasing rate
  - EOSDIS archive just exceeded one petabyte in February of this year
  - New missions (e.g., Aqua) will put additional strains on existing archive services or require additional services
  - User interfacing and data selection are a challenge due to increasing volumes of data and the distributed nature of archives
- Space Science's virtual observatory archiving is expected to be as demanding as Earth science's in the near future
  - Virtual observatory's data volumes will match Earth science's as the program matures
  - Data sources and archives will be distributed (expected to be located close to land based observing instruments)
- The Intelligent Archive Project is formulating strategies and architectures to help resolve the challenges in archiving for Earth and space sciences that result from
- Ever increasing amount of data volumes and rates
  - Increasing numbers of missions and data sources
  - Increasing demand to support greater numbers of scientists and areas of research
  - Heterogeneous and distributed environment of data providers/users
  - Complexity of data, information and knowledge

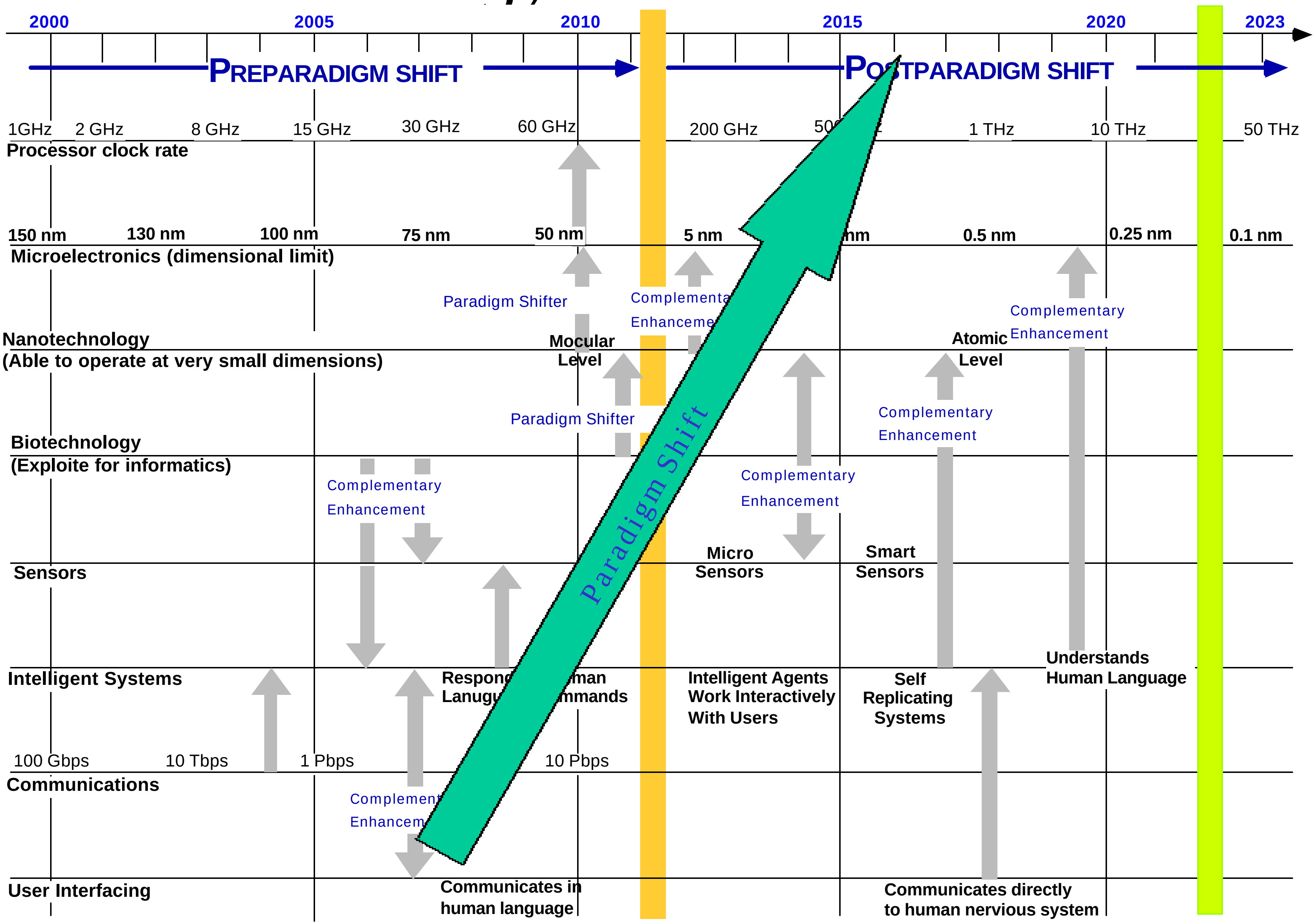
Preliminary Findings

- Future missions will significantly increase the amount and kind of data to be archived and managed
- Expect large numbers of distributed users with personal computing and storage resources that could become a source of data to be archived
- Functions that will need automation and infusion of intelligence
  - Data acquisition, cataloging and characterization
  - Production operations
  - Transformation of data into information
  - User (human and computer) access and communication
  - Forecast and prediction model support
  - Storage and supporting management strategies
  - System management, communications and planning
- Science missions will commonly include models and simulations
- Modeling systems may become an “archive user” that will task sensors, in near real time, to collect data to support simulation analysis,
- Models could request sensors for specific acquisitions of data
  - Requested data will need to be processed in a timely manner
  - The number of sensors that could be tasked may be large in number, and distributed

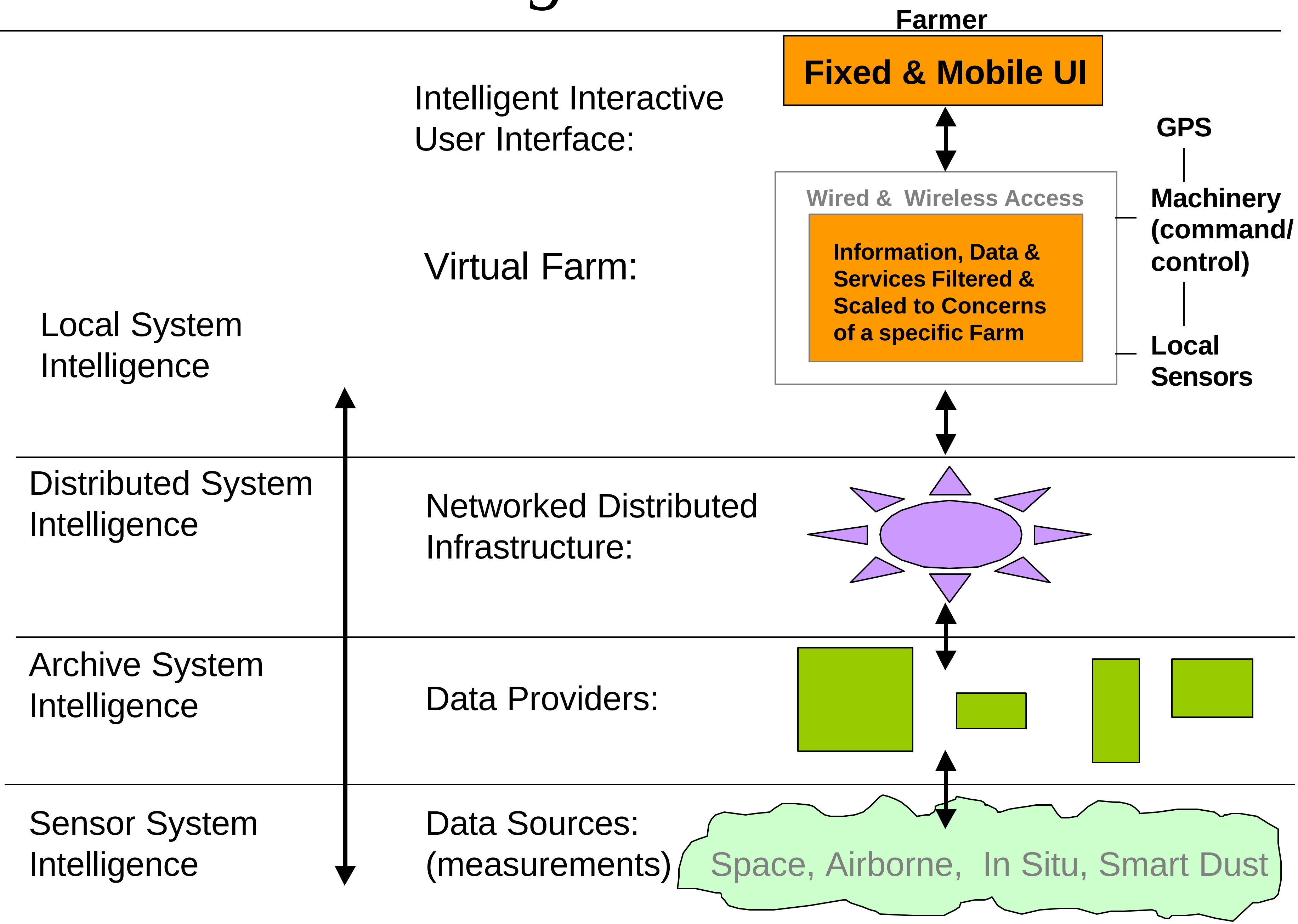
Preliminary Findings (Cont.)

- Use cases have provided valuable information related to archive functionality, scope, and performance
  - Two use cases have been evaluated; precision farming and weather
- Assessed and characterized technology evolution over next 20 years
  - Expect a major paradigm shift (See diagram next viewgraph) which will have a fundamental impact on functionality and performance
- Formulated architectures that relate technologies to core functions
- Topology and texture of architecture very likely to be adaptable and evolutionary
- Archives may store only limited levels of data and produce virtual data products on-demand
- Existing archive systems will need to be integrated into a future intelligent archive
- Intelligent archive will utilize distributed computing to extent possible

Technology Forecast Timeline



Precision Agriculture Scenario



Precision Agriculture Support Information

Data Volumes

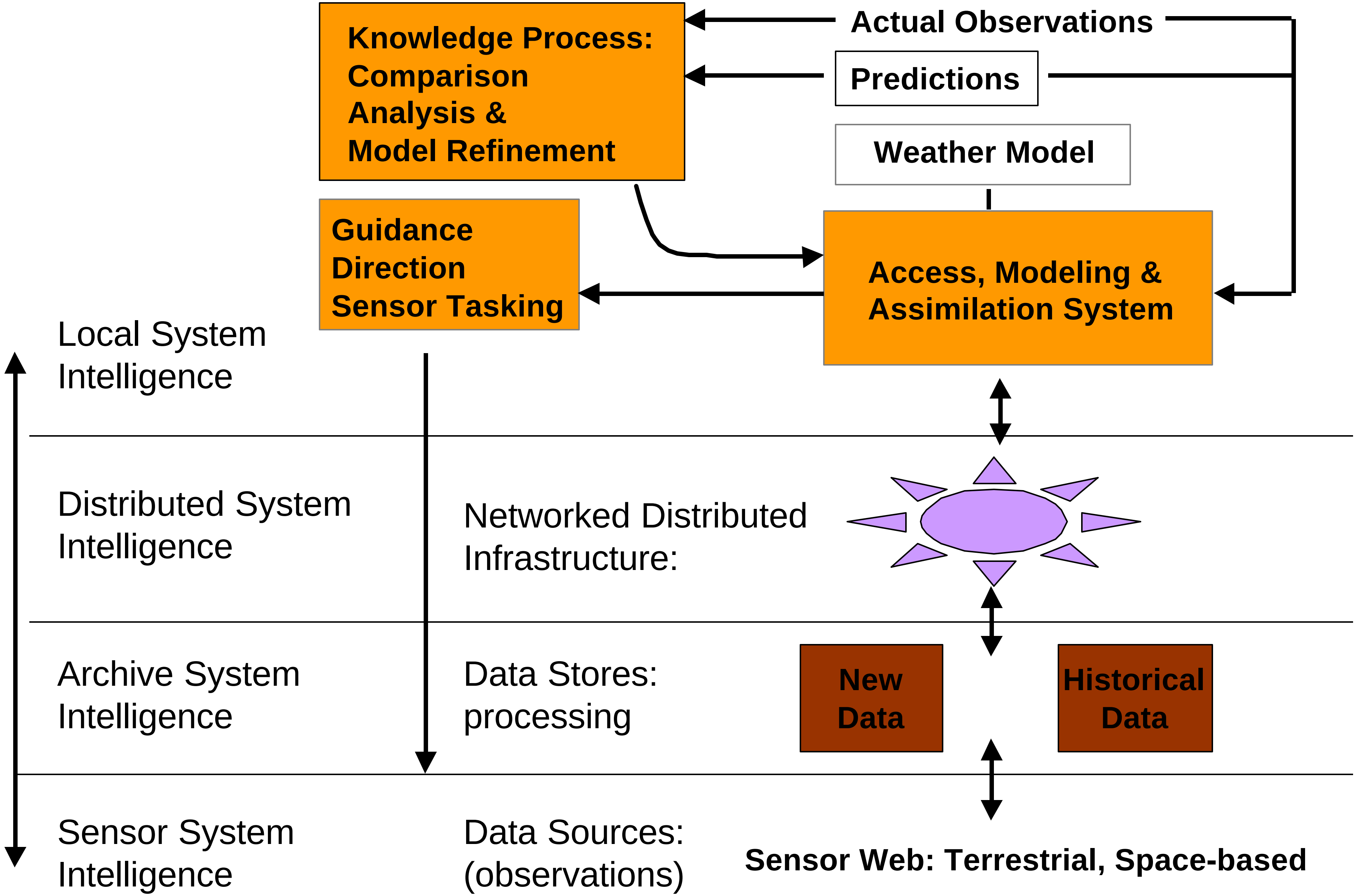
- Estimated 1.5 TB per year for a 1000-acre farm, (including satellite and airborne remote sensing data, in situ data, visualizations, modeling, etc.)
- As of 2001 there were 2,158,000 farms in the U.S. (averaging 436 acres each)*
- Total acreage, 9.4×10^{8}. Result in data stream of 1.4 EB per year for all U.S. farms.

Data Sources

- Remote sensing at spatial resolutions as small as 1 foot and temporal resolutions as fine as 1 hour;
- Precision weather forecasts from three hours ahead to longer-range climate predictions for months and years;
- Land profiles including soils, moisture, elevations, drainage patterns and water sources, digital ortho-photo quadrangles (images with integrated USGS topography maps), digital elevation models, geo-rectified spatial data, ecological zone profiles, biodiversity inventory, local calibrations including ground truth.
- Types of analysis needed
- Planting conditions indicators, crop monitoring (growth, maturity, health and stress indicators),
- Weed and pest identification and tracking, chemical and other intervention impact prediction and analysis,
- Weather conditions, advance forecasts, environmental alerts, microclimate surveys, soil types and depths.

* Source: Wisconsin Agricultural Statistics Service via USDA

Precision Weather Scenario



Precision Weather Support Information

Data Volumes

- Approximately 7.5 TB/day by 2025*

Data Sources

- Space-based, airborne and terrestrial sensors networked via fiber and wireless,
- Large forecast models (forecast models will be approaching 100 million unique grid points with a million observations

Unique Features

- Modeling systems directly task sensors
- Large number of sensors
- Archive is an integral part of the science/modeling process

Types of analysis required

- Structure information in the free atmosphere every 3 hours, every 25 km globally, and vertically from the surface to 80 km altitude
- Global 3D distribution of cloud height, cloud depth, aerosols, water/ice, and suspended precipitation rates
- Land and sea surface temperature, land surface moisture, albedo, vegetation type
- Planetary boundary layer depth

* *Source: Advanced Weather Prediction Technology: NASA's Contribution to the Operational Agencies, Vision 2025 Architecture Study*

Accomplishments

- Selected and characterized two science/application scenarios - precision weather and precision farming
- Formulated an abstract functional architecture and two conceptual physical architectures
- Studied GSFC DAAC user interactions and demands, proof of concept of knowledge feedback and its relationship to the data from which it was derived
- Presented paper at IEEE Mass Store Conference (College Park, MD, April 2002)
- Submitted paper to "Future Intelligent Earth Observing Systems" (FIEOS) conference (Denver CO, November 2002)
- Prepared a preliminary report on work to date
- Identified additional technical issues critical to study objectives on which "drill-down" white papers will be developed

Technical Significance of Progress & Expected Impact

Significance of Progress

- Have formulated a mission-relevant context for the inclusion of IDU technologies
- Have conceptualized architectures that will support IDU technologies
- Assessed technology progress in the 2015-2020 time frame and identified their impact on future IDU and IA systems

Expected Impact

- Future IA systems will be able to deal with the large data volumes and rates expected from future missions
- Defined systems that are able to support intelligent processes that extract information content from spatial data
- Defining a roadmap that will support the utilization of IDU research in the implementation of future mission archives

Technical Issues &Risks

Mission Drivers

- The success of future science missions will become increasingly dependent on data archiving services and how well information can be automatically extracted from data
- Future missions can be expected to use large numbers of sensors which will significantly increase the amount and kind of data to be archived and managed
- Data volumes continue to increase rapidly

Automated Analysis and Understanding

- Data mining is not yet able to function at human levels of performance for the identification and classification of features and phenomena
- Machine learning is only marginally successful in acquiring information and knowledge from humans
- Ability of IDU data mining algorithms to perform better than human experts remains untested
- Commercial market tools are limited in handling complex science data

Technical Issues &Risks (Cont.)

Archiving Technologies

- Changes in storage technologies will continue to force constant refreshing of data in archives
  - As the data volumes increase, this will become a very costly long-term processing issue
- Retiring aged systems will be a problem (they have mortality)
- Full automation of archiving function is difficult to achieve
  - Science data complexity often forces significant manual intervention.
  - Standardization of data models would help, but move is toward increasing heterogeneity since populating data models is burdensome
  - Many failures stem from software errors, which are resistant to automated fail-over

URL Links To Research Activities

The link to the Intelligent Archive Project web site:

http://daac.gsfc.nasa.gov/IDA/

References

- H. K. Ramapriyan, S. Kempler, C. Lynnes, G. McConaughy, K. McDonald, R. Kiang, S. Calvo, R. Harberts, L. Roelofs, D. Sun, “Conceptual Study of Intelligent Data Archives of the Future”, *10th NASA Goddard Conference on Mass Storage Systems and Technologies and 19th IEEE Symposium on Mass Storage Systems*, April 15-18, 2002, College Park, MD. *http://storageconference.com/2002/index.html*
- D. Sun, C. Lynnes, R. K. Kiang, S. Kempler, G. Serafino, “Knowledge discovery about scientific papers or proceedings referenced NASA/DAAC data with a rule-based classifier”, *SPIE Conference*, *Proceedings of SPIE Vol. #4730*, April 1-5,2002, Orlando, FL. *http://www.spie.org/Conferences/Programs/02/or/confs/4730.html*